ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 8, Issue 4, April 2021

Action Identification of Humans in 2D Images Using Transfer Learning

Sai Siddharth Upadhyayula[1], Nidhish Sharma[2], Ipsita Sahu[3], Vineet Gandhi[4], Dasari Prasad[5]

UG Scholar, Department of Computer Science and Engineering, GITAM Institute of Technology, GITAM (Deemed to be University), Visakhapatnam, Andhra Pradesh, India[1,2,3,4]

Assistant Professor, Department of Computer Science and Engineering, GITAM Institute of Technology, GITAM (Deemed to be University), Visakhapatnam, Andhra Pradesh, India[5]

Abstract: Human Activity Recognition (HAR) is a well-studied and vital topic in Computer Vision. Action Recognition includes many essential applications such as security surveillance, healthcare, and pedestrian monitoring for autonomous vehicles. In recent years with advancements in both computing power and research in Deep Learning, these applications have made many breakthroughs. Here for recognizing human actions in still images we make use of a multi-cue-based approach, where the pertinent object areas are discovered and used in a weakly supervised manner. We do not require an explicitly trained object detector or part/attribute annotation. A Semantic Segmentation approach is used over sets of objects hypotheses in order to represent objects related to the actions. We test method on the extensive Stanford 40 Actions dataset and achieve significant performance gain. The results we obtained showcase that using multiple object hypotheses within Semantic Segmentation is effective for human action recognition in still images and representation of object is suitable for using in conjunction with other visual features. The transfer learning technique was implemented on the Stanford 40 dataset by making use of CNN-based architectures like Inception V-3, Resnet-50, VGG-16 and Efficient Net B7. Of all the CNN-based architectures tested on the Stanford 40 dataset, Efficient Net produced the highest accuracy with 67.43%.w

Keywords: **Human Activity Recognition (HAR), Transfer learning, CNN, Efficient Net B7.**

I. INTRODUCTION

Perceiving activities in still pictures has as of late acquired consideration in the vision local area because of its enormous pertinence to different areas. In news photos, for instance, it is particularly critical to comprehend what individuals are doing from a recovery perspective. Instead of movement and appearance in recordings, actually pictures pass on the activity data by means of the posture of the individual and the encompassing article/scene setting. Items are particularly significant signals for distinguishing the sort of the activity.

Here we approach the issue of recognizing related items from a pitifully administered perspective and investigate the impact of utilizing Transfer Learning for discovering the competitor object locales and their comparing impact in acknowledgment. Move Learning utilizes a pre-prepared model. This model has been prepared on an incredibly huge dataset, and we would have the option to move loads which were learned through many long periods of preparing on numerous powerful GPUs.

Numerous such models are publicly released, for example, VGG-19 and Inception-v3. They were prepared on great many pictures with incredibly high registering power which can be pricey to accomplish without any preparation. Move Learning has become enormously mainstream since it impressively decreases preparing time, and requires much less information to prepare on to expand execution.

Advantages and Disadvantages

The principle benefits of HAR frameworks are to notice and break down human exercises and to decipher continuous occasions effectively. Utilizing visual and non-visual tactile information, HAR frameworks recover and measure context oriented (natural, spatial, transient, and so forth) information to comprehend the human conduct. There are a few application spaces where HAR ideas are researched and the frameworks are created. We partition them generally into four classes: dynamic and helped living (AAL) frameworks for

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 8, Issue 4, April 2021

shrewd homes, medical services observing applications, checking and reconnaissance frameworks for indoor and open air exercises, and tele-drenching (TI) applications.

Customarily, the errand of noticing and dissecting human exercises was done by human administrators, for instance, in security and reconnaissance measures or the cycles of checking a patients' ailment. With the expanding number of camera perspectives and specialized observing gadgets, in any case, this undertaking gets more trying for the administrators as well as progressively cost-serious, specifically, since it demands nonstop activity. Also, HAR frameworks inside these fields can uphold or even supplant human administrators to upgrade the productivity and viability of the perception and investigation measure.

Perceiving human exercises from video groupings or still pictures is a difficult errand because of issues, for example, foundation mess, halfway impediment, changes in scale, perspective, lighting, and appearance. Numerous applications, including video reconnaissance frameworks, human-PC cooperation, and mechanical technology for human conduct portrayal, require a various movement acknowledgment framework.

II. LITERATURE SURVEY

On Recognizing Actions in Still Images via Multiple Features by Fadime Sener, Cagdas Bas, and Nazli Ikizler-Cinbis **[1]**

Key points:

The paper gave a deeper intuition into Multi Instance Learning (MIL)

The paper also provided information regarding how multiple features are deduced from images in various cases

Improvements:

The previous strategy in Human Activity Recognition depends on part and quality portrayal, where each picture is addressed through an inadequate arrangement of "activity bases".

The strategy proposed in the paper proposed another technique called Multiple Instance Learning (MIL) which fundamentally expanded the exhibition of the model.

Human Action Recognition by Learning Bases of Action Attributes and Parts by Bangpeng Yao, Xiaoye Jiang, Aditya Khosla, Andy Lai Lin, Leonidas Guibas, and Li Fei-Fei **[2]**

Key points:

The paper proposed an attribute and parts-based representation of human actions in a weakly supervised setting.

We define these connections of activity credits and parts as activity bases for communicating human activities. A specific activity in a picture can hence be addressed as a weighted summation of a subset of these bases.

The paper showed how Stanford 40 dataset is better compared to other datasets in the field of Human Activity Recognition in still pictures.

Improvements:

The previous technique in Human Action Recognition utilized the entire picture to address an activity and treat activity acknowledgment as an overall picture characterization issue.

These strategies don't, in any case, investigate the semantically significant segments of an activity, for example, human postures and the articles that are firmly identified with the activity.

The current strategy proposed in the paper assisted with understanding the activities addressed in the picture in a superior way.

Activity Recognition with Still Images and Video by Wendell Hom **[3]**

Key points:

The paper proposed another technique called Loss Guided Act for the Human Action Classification in still pictures and recordings.

For still pictures, the paper utilized exchange learning on various models pre-prepared on ImageNet loads, supplanting the last layer with a 40-way SoftMax yield layer. The CNN base layers were frozen.

The paper executed different designs in the pictures utilizing Stanford 40 dataset which assisted us with picking the correct engineering for the venture.

Improvements:

The previous techniques in the field of Human Action Recognition either utilized Semantic body examination or thinking about the whole picture for the arrangement of different activities.

The paper proposed another technique called Loss Guided Act.

Still Image-based Human Activity Recognition with Deep Representations and Residual Learning by Ahsan Raza Siyal, Zuhaibuddin Bhutto, Syed Muhammad Shehram Shah, Azhar Iqbal4, Faraz Mehmood, Ayaz Hussain and Saleem Ahmed **[4]**

Key points:

The design utilized in the paper is a pretrained CNN (Convolutional Neural Network) trailed by a SVM (Support Vector Machine).

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 8, Issue 4, April 2021

The CNN is utilized as an element extractor and SVM is utilized for activity acknowledgment.

The proposed technique is assessed on openly accessible stanford40 human activity informational index, which incorporates 40 classes of activities and 9532 pictures.

Improvements:

The proposed strategy accomplishes better execution over traditional techniques in term of exactness and computational force.

***A Survey on Still Image Based Human Action Recognition by Guodong Guo and Alice Lai* [5]**

Key points:

The paper is a review of the existing approaches to still image-based action recognition. Various high-level cues and low-level features for action analysis in still images are categorized and described.

III. SYSTEM METHODOLOGY

A. Transfer Learning

In transfer learning, the information on an all-around prepared AI model is applied to an alternate yet related issue. For instance, on the off chance that you prepared a straightforward classifier to anticipate whether a picture contains a rucksack, you could utilize the information that the model acquired during its preparation to perceive different items like shades.

With transfer learning, we essentially attempt to misuse what has been realized in one errand to improve speculation in another. We move the loads that an organization has learned at "task A" to another "task B."

B. How it functions?

In PC vision, for instance, neural organizations normally attempt to distinguish edges in the prior layers, shapes in the center layer and some assignment explicit highlights in the later layers. In move learning, the early and center layers are utilized and we just retrain the last layers. It helps influence the named information of the undertaking it was at first prepared on.

In move learning, we attempt to move however much information as could be expected from the past task the model was prepared on to the new job needing to be done. This information can be in different structures relying upon the issue and the information. For instance, it very well may be the way models are formed, which permits us to all the more effectively distinguish novel articles.

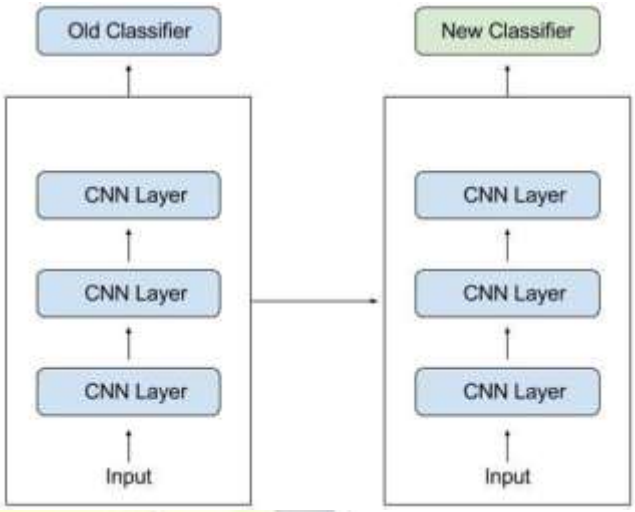


Fig. 1 Transfer Learning Methodology

C. Efficient Net B7 Architecture

Proficient Net is quickly supplanting ResNet as the foundation of decision for some, profound learning undertakings. It's at the highest point of the ImageNet arrangement leader boards.

The thought: Many of the models in this blog entry have a "multiplier" or some likeness thereof that allows you to decide the number of channels each layer has. A more modest multiplier makes the model quicker, a bigger multiplier makes the model more precise. However, that is by all account not the only method to scale a convnet. There are really three unique things that can be increased or down:

Organization width: what number channels there are per convolution layer. This is the one I just referenced.

Organization profundity: what number layers there are in each phase of the model. The quantity of stages remains something very similar, however stages can be made further by adding more layers to them.

Info goal: The components of the information picture. It's not unexpected to utilize 224×224 pictures, yet making the goal higher will by and large improve the precision (to a certain degree).

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 8, Issue 4, April 2021

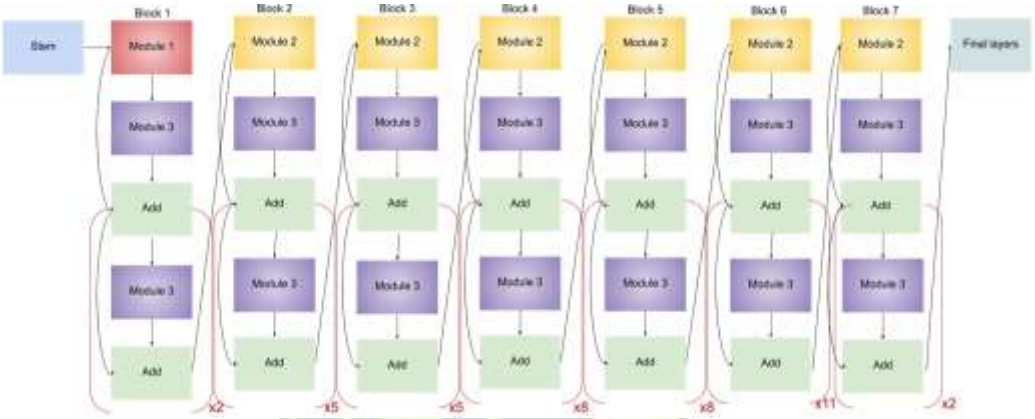


Fig. 2 Efficient Net B7 Model

The baseline model in the Efficient Net architecture is the B0 variant and it is updated till the B7 variant.

IV. SYSTEM IMPLEMENTATION

A. Dataset

The Stanford 40 Action Dataset contains pictures of people performing 40 activities. In each picture, we give a jumping box of the individual who is playing out the activity showed by the filename of the picture. There are 9532 pictures altogether with 180-300 pictures for every activity class.

TABLE 1: The Stanford 40 Actions dataset: the list of actions and number of images in each action

Action Name	# imgs	Action Name	# imgs
Applauding	279	Reading book	234
Blowing bubbles	292	Repairing a car	122
Brushing teeth	211	Riding a bike	288
Calling	272	Riding a horse	260
Cooking	295	Rowing a boat	149
Cutting trees	175	Running	254
Cutting vegetables	131	Shooting an arrow	211
Drinking	194	Smoking cigarette	175
Feeding a horse	319	Taking photos	154
Fishing	269	Throwing a frisby	195
Fixing a bike	131	Using a computer	230
Filling up gas	123	Using a microscope	127
Hanging clothes	121	Using a telescope	151
Holding an umbrella	289	Using an ATM	144
Jumping	299	Walking a dog	294
Mopping the floor	159	Washing dishes	183
Playing guitar	295	Watching TV	146

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

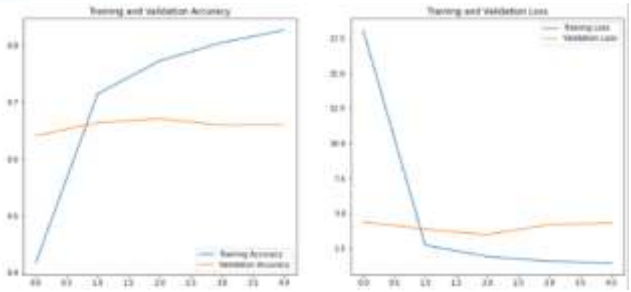
International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 8, Issue 4, April 2021

Playing violin	268	Waving hands	209
Poling a boat	118	Writing on a board	133
Pushing a cart	172	Writing on a book	127

Fig. 3 Example images of the Stanford 40 Actions Dataset

B. Plotting the accuracy and loss graphs

Now the graphs for accuracy and loss are mapped



ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 8, Issue 4, April 2021

Fig. 4 Accuracy and Loss of Inception V3

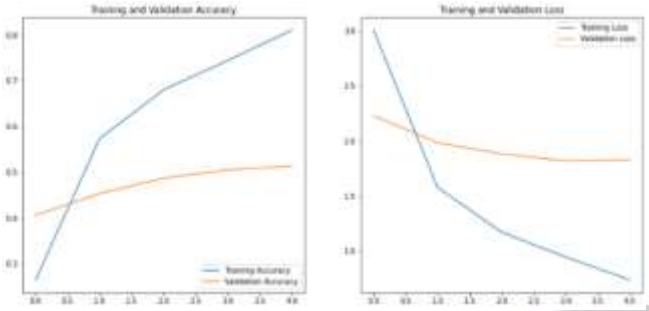


Fig. 5 Accuracy and Loss of VGG 16

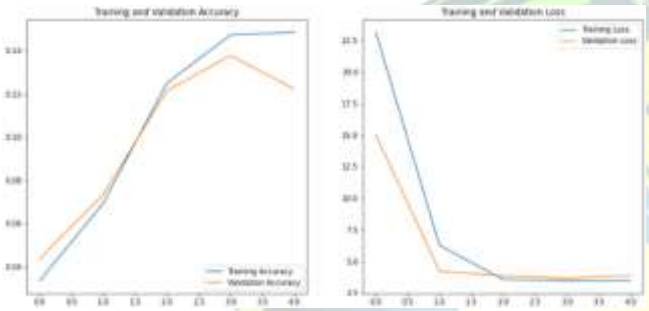


Fig. 6 Accuracy and Loss of Resnet 50

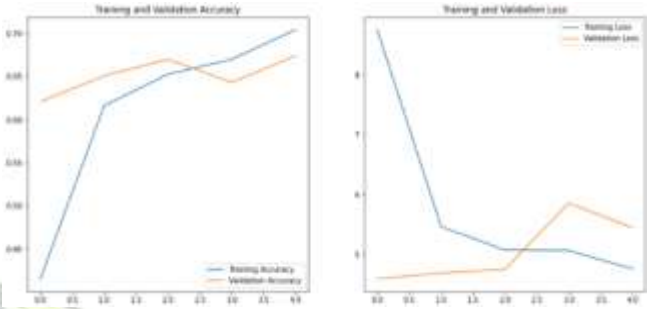


Fig. 7 Accuracy and Loss of Efficient Net B7

V. TESTING

A.Testing on Efficient Net B7 Architecture

	precision	recall	f1-score	support
applauding	0.39	0.51	0.44	100
blowing_bubbles	0.79	0.49	0.60	100
brushing_teeth	0.52	0.55	0.53	100
cleaning_the_floor	0.90	0.87	0.88	100
climbing	0.81	0.92	0.86	100
cooking	0.41	0.64	0.50	100
cutting_trees	0.96	0.79	0.87	100
cutting_vegetables	0.59	0.55	0.57	100
drinking	0.43	0.37	0.40	100
feeding_a_horse	0.94	0.88	0.91	100
fishing	0.69	0.90	0.78	100
fixing_a_bike	0.59	0.94	0.72	100
fixing_a_car	0.94	0.73	0.82	100
gardening	0.86	0.65	0.74	100
holding_an_umbrella	0.87	0.95	0.91	100
jumping	0.82	0.81	0.81	100
looking_through_a_microscope	0.74	0.73	0.74	100
looking_through_a_telescope	0.78	0.73	0.76	100
phoning	0.39	0.22	0.28	100

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 8, Issue 4, April 2021

playing_guitar	0.86	0.98	0.92	100
playing_violin	0.89	0.88	0.88	100
pouring_liquid	0.44	0.40	0.42	100
pushing_a_cart	0.89	0.64	0.74	100
reading	0.63	0.19	0.29	100
riding_a_bike	0.92	0.70	0.80	100
riding_a_horse	0.93	0.94	0.94	100
rowing_a_boat	0.95	0.97	0.96	100
running	0.66	0.78	0.71	100
shooting_an_arrow	0.78	0.97	0.86	100
smoking	0.27	0.44	0.33	100
taking_photos	0.52	0.32	0.40	100
texting_message	0.30	0.30	0.30	100
throwing_frisby	0.92	0.68	0.78	100
using_a_computer	0.68	0.68	0.68	100
walking_the_dog	0.86	0.85	0.85	100
washing_dishes	0.51	0.42	0.46	100
watching_TV	0.76	0.89	0.82	100
waving_hands	0.40	0.44	0.42	100
writing_on_a_board	0.80	0.56	0.66	100
writing_on_a_book	0.43	0.71	0.53	100
accuracy			0.67	4000
macro avg	0.69	0.67	0.67	4000
weighted avg	0.69	0.67	0.67	4000

VI. RESULTS

A. Overall Accuracy

Table 2: Comparison of Testing Accuracy

ALGORITHM	TRAINING ACCURACY	TRAINING LOSS	TESTING ACCURACY	TESTING LOSS
INCEPTION V3	82.24	1.5648	66.03	4.3219
VGG 16	81.68	0.7247	51.35	1.8290
RESNET 50	15.22	3.3776	12.22	3.8927
EFFICIENT NET B7	72.42	4.3085	67.43	5.4496

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 8, Issue 4, April 2021

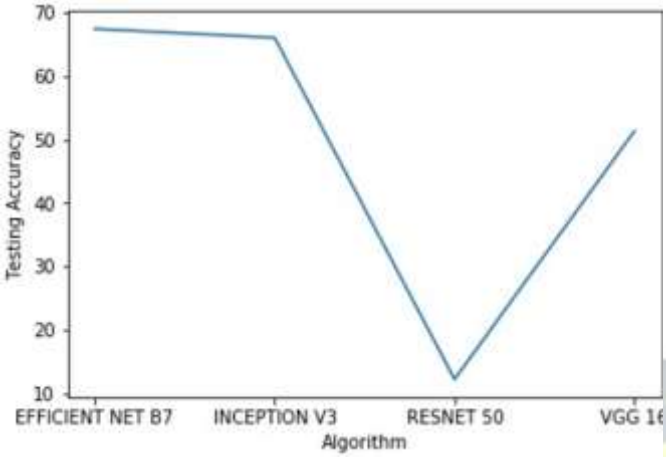


Fig. 8 Comparison of Testing Accuracy

VII. CONCLUSION

In the wake of testing the transfer learning procedure with the semantic division technique by utilizing four CNN-based models like Inception – V3 architecture, Resnet – 50 architecture, VGG – 16 architecture and Efficient net B7 architecture on the Stanford 40 dataset, we presume that the **Efficient net B7 architecture** is the best design for recognizing activities of people among the tried structures inferable from the way that it has the most elevated testing precision of **67.43** and it has a superior parted of the preparation and testing accuracy.

REFERENCES

[1]. Sener F., Bas C., Ikizler-Cinbis N. (2012) On Recognizing Actions in Still Images via Multiple Features. In: Fusiello A., Murino V., Cucchiara R. (eds) Computer Vision – ECCV 2012. Workshops and Demonstrations. ECCV 2012. Lecture Notes in Computer Science, vol 7585. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-33885-4_27

[2]. Yao, Bangpeng & Jiang, Xiaoye & Khosla, Aditya & Lin, Andy & Guibas, Leonidas & Li, Fei-Fei. (2011). Human action recognition by learning bases of action attributes and parts. ICCV. 1331-1338. 10.1109/ICCV.2011.6126386.

[3]. Wendell Hom., Activity Recognition with Still Images and Video, CS230: Deep Learning, Winter 2020, Stanford University, CA

[4]. Siyal, Ahsan & Bhutto, Zuhaibuddin & Syed, Muhammad Shehram Shah & Iqbal, Azhar & Mehmood, Faraz & Hussain, Ayaz & Ahmed, Saleem. (2020). Still Image-based Human Activity Recognition with Deep Representations and Residual Learning. International Journal of Advanced Computer Science and Applications. 11. 471-477. 10.14569/IJACSA.2020.0110561.

[5]. Guo, Guodong & Lai, Alice. (2014). A survey on still image based human action recognition. Pattern Recognition. 47. 3343–3361. 10.1016/j.patcog.2014.04.018.

[6]. VGG16 – Convolutional Network for Classification and Detection by Muneeb ul Hassan

[7]. Accelerating Very Deep Convolutional Networks for Classification and Detection by Xiangyu Zhang, Jianhua Zou, Kaiming He and Jian Sun

[8]. Understanding and Coding a ResNet in Keras

[9]. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks by Mingxing Tan and Quoc V. Le

[10]. Human Action Recognition by Learning Bases of Action Attributes and Parts. *Internation Conference on Computer Vision (ICCV)*, Barcelona, Spain. November 6-13, 2011 by B. Yao, X. Jiang, A. Khosla, A.L. Lin, L.J. Guibas, and L. Fei-Fei.

[11]. UNDERSTANDING HUMAN ACTIONS IN STILL IMAGES by Bangpeng Yao